REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

AD-A242 933

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
AUG 91	Technical Report 01 JUL 87 - 30 SEP 90

4. TITLE AND SUBTITLE
OPTIMAL SEQUENTIAL DESIGNS FOR ITEM ESTIMATION (UNCLASSIFIED)

5. FUNDING NUMBERS
PE NO. 0602233N
PR NO. RM33M20

6. AUTHOR(S)
DOUGLAS H. JONES

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
THATCHER JONES ASSOCIATES
1280 WOODFERN COURT
TOMS RIVER, NJ 08648

8. PERFORMING ORGANIZATION REPORT NUMBER
91-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
COGNITIVE SCIENCE PROGRAM
OFFICE OF NAVAL RESEARCH (CODE 1142CS)
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NOOO14 -87-C-0696,
R&T 4428007

11. SUPPLEMENTARY NOTES
SUPPORTED BY THE OFFICE OF THE CHIEF OF NAVAL RESEARCH MANPOWER, PERSONNEL, AND TRAINING R&D PROGRAM.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Replenishing item pools for on-line ability testing requires innovative and efficient data collection designs. Based on a theoretical framework for generating exact D-optimal-designs for selecting individual examinees, and for consistently estimating item parameters, this article presents a sequential procedure for on-line item calibration. These procedures were derived for general, dichotomous item response models, using Welch (1982) for exact n-point D-optimal designs and Stefanski and Carroll (1985) for consistent estimators. In simulations, these designs appear to be considerably more efficient than random seeding of items.

91-16372

14. SUBJECT TERMS
On-Line Testing, Sequential Design, Integer Programing, Item Response Theory, Computerized Adaptive Testing, Exact n-Point D-Optimal Designs, Measurement Errors, On-Line Item Calibration

15. NUMBER OF PAGES

16. PRICE CODE

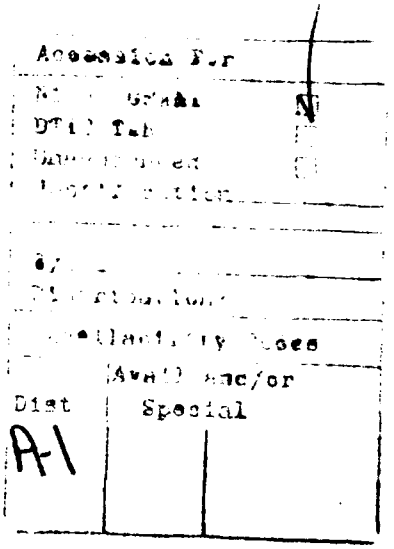
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

91 1122 104

Optimal Sequential Designs for On-Line Item Estimation[1] (Unclassified)

Douglas H. Jones[2]
Graduate School of Management
Rutgers, The State University of New Jersey

30 Aug 1991



---

[1]This report was prepared under the Navy Manpower, Personnel, and Training R&D Program of the Office of the Chief of Naval Research under Contract N00014-87-C-0696.

[2] Address all correspondences to Douglas H. Jones, Rutgers, The State University, Graduate School of Mangement, 92 New Street, Newark, NJ 07102. The author wishes to acknowledge the advice and consultation given by Ronald Armstrong, Charles Davis, Bradford Sympson, Zhabao Wang and In-Long Wu. Special programming assistance was provided by Mr. Zhi-Yang Jin.

Optimal Sequential Designs for On-Line Item Estimation

Executive Summary

The workhorses of modern test theory are so-called item characteristic curves (ICC's); these are mathematical functions which describe how the probability of correctly answering a test question changes with ability. In CAT testing, these ICC's are used both to select appropriate problems for an examinee and to score the examinee's performance on the selected problems.

The ICC's for a test's problems are not known, but must be estimated from data. Typically, such data are collected before the test is used operationally in what is known as a "calibration study". In the case of CAT-ASVAB, calibration data were collected by administering subsets of the questions via paper-and-pencil test booklets to 50,000 applicants for military service in Military Entrance Processing Stations (MEPS).

Such off-line calibration studies have several shortcomings: First, they are expensive to conduct in that they make heavy demands on an overburdened MEPS command and sometimes prevent same-day processing, necessitating that applicants be billeted in hotels. Second, the performance data are suspect, since examinees are told that their performance on the non-operational problems will not "count". Third, the process is inefficient in that the random sample of examinees given a particular test question, is usually not the optimal sample for estimating that question's ICC.

It is widely held that the answer to the shortcomings of "off-line" calibration is "on-line" calibration. In on-line calibration one gathers the

needed data to estimate ICC's by unobtrusively seeding a small number of non-operational items into an applicant's operational CAT test. If the number of additional items given to each applicant is small, data collection is virtually cost free. If an applicant cannot distinguish non-operational items from operational items, the performance data will better reflect his/her capabilities. And, if the non-operational items are embedded in an operational test which is administered via computer, in principle one should not be limited to collecting data from random samples, but could employ some optimal sample design strategy. This work seeks to develop the wherewithal for dynamically constructing optimal samples.

One can think of the optimal sample design problem in the following terms. For concreteness suppose that we have 500 new test problems to be calibrated and that each day 1000 applicants are tested in the MEPS. Further suppose that we can tolerate at most 2 non-operational items embedded in each applicant's operational test. Since each applicant can be assigned 2 items from a set of 500 items, there are 500 choose 2, or 124,750, potential allocations for each applicant. Since there are 1000 applicants, the number of potential allocations overall is astronomical (or to be exact 1.2475E1004, where xEn means x multiplied by 10 to the power n). The sample design problem is the problem of allocating the available applicants to the non-operational items in an optimal manner.

If one is to improve upon random allocation, one needs three elements: (a) a relevant basis on which to distinguish the objects to be allocated (in this case the applicants), (b) an objective function which orders the set of potential allocations with respect to some measure of

quality, and (c) an efficient algorithm for "searching" the rather large space of potential allocations.

This research developed all three elements. (a) Since examinee abilities are not known, the applicants were distinguished by the maximum-likelihood estimates of their ability from their operational CAT test. (b) The objective function was based on the determinant of the Fisher information matrix for the parameters of the ICC. And, (c) a branch and bound algorithm was used to search the space of potential allocations for the set which is optimal.

The need for several approximations accompany adoption of an objective function based on the Fisher information matrix. First, to compute the Fisher information matrix, one must know the ICC. This requirement was circumvented by employing a sequential optimization strategy in which initial ICC estimates were iteratively refined as more and more data were gathered. The method developed for updating ICC estimates involved modeling the measurement error of the CAT ability estimate and using this model to modify the maximum likelihood estimate of the item parameters. Without this modification the usual MLE is biased.

Optimal allocation via Fisher information also requires that abilities be known. In this case maximum-likelihood estimates of ability were used. The maximum-likelihood estimates are based on the data from the operational portion of the CAT.

Monte Carlo data suggest that the consequences of these approximations were not severe. Even without knowledge of the actual ICC's or abilities, the approach proved to be at least 90% as efficient as the theoretically optimal design in all cases studied. Designs based on random

allocation of applicants to items were less than 30% as efficient as the sequential design algorithm developed here.

Optimal Sequential Designs for On-Line Item Estimation

Abstract

Replenishing item pools for on-line ability testing requires innovative and efficient data collection designs. Based on a theoretical framework for generating exact D-optimal-designs for selecting individual examinees, and for consistently estimating item parameters, this article presents a sequential procedure for on-line item calibration. These procedures were derived for general, dichotomous item response models, using Welch (1982) for exact n-point D-optimal designs and Stefanski and Carroll (1985) for consistent estimators. In simulations, these designs appear to be considerably more efficient than random seeding of items. *Key words: Branch-and-bound, Computerized adaptive test, Exact n-point D-optimal, Integer programming, Item response theory, Measurement errors model, On-line testing, Sequential design.*

Introduction

Calibration of new items is an essential part of a testing system, because operational items eventually become overexposed and need replacement. To calibrate new items for the Armed Services Vocational Aptitude Battery (ASVAB), costly testing sessions are conducted where all the new items are presented to examinees that have been recruited expressly for the purpose (C. E. Davis, personal communication, March 15, 1991). The obtained data may be unreliable because the examinees know that the test "doesn't count" and, thus, do not do their best. On-line item calibration promises to yield more reliable data on new items at virtually no cost. This research is concerned with the development of item calibration procedures that take advantage of the auxiliary ability estimates supplied by the on-line test. This will enable the procedure to select pre-specified ability distributions known to yield high information regarding a given item.

Researchers have recently focussed on the effect of an ability distribution on the precision of the estimate of an item parameter. Wingersky and Lord (1984) showed that when item and ability parameters are estimated simultaneously, a rectangular distribution of ability, instead of a normal distribution, reduces the standard errors of all parameters. Studying the standard errors of the estimates of the item parameters only, Stocking (1990) concluded that a broad distribution of ability, uniform or bimodal, was better than a bell-shaped distribution. In addition, she convincingly argued that even in very large samples, very little information may be available for calibrating some items, and that the success of a

particular item calibration using item response theory depends heavily on the selection of more informative data.

The theory of optimal design is concerned with planning data collection so that they will be as informative as possible. The general theory of optimal design focuses on "minimizing" the variance-covariance matrix of the parameter estimates, or on "maximizing" the inverse of the Fisher information matrix. A D-optimal design maximizes the determinant of the Fisher information matrix; and, an A-optimal design minimizes the trace of the inverse of the Fisher information matrix (Federov, 1972; and Silvey, 1980). The criteria used in Wingersky and Lord (1984) and Stocking (1990) are related to the theory of A-optimality.

Ford (1976) derives D-optimal designs for logistic regression functions, also known as two-parameter logistic item response models (Lord, 1980). He shows that discrete, two-point distributions are D-optimal, with support points depending on the values of the item parameters. However, optimal designs for the two-parameter logistic functions are unstable. Silvey (1980) gives an example showing that Ford's design, optimal for one item, may be extremely suboptimal for another item, even if its parameters are close-by. Silvey concludes that is not practical to use a design that is maximin over a subset of candidate values for the unknown parameters. A design is maximin if it maximizes the minimum possible determinate of the Fisher information matrix. To overcome this problem, Ford and Silvey (1980) study sequentially constructed designs for two-parameter logistic models, which employ D-optimal "subdesigns" based on current estimates of the parameters. This research will show that sequentially constructed designs are useful for on-line item calibration.

Ford and Silvey's sequentially constructed designs apply large sample optimal design theory to small sample subdesigns. The approximation of small exact designs with large sample designs is sometimes inadequate (Welch, 1982). Recently, researchers have developed some theory and procedures for finding optimal designs for small samples, called optimal exact N-point designs. For linear regression models, Welch (1982) investigated a branch-and-bound algorithm for finding D-optimal exact N-point designs with support over discrete design space. Also for linear regression models, Haines (1985) investigated a simulated annealing algorithm for N-point designs with support over a continuous design space. Additional algorithms are presented in Donev and Atkinson (1988) and Mitchell (1974). This research will find exact N-point designs for constructing on-line calibration samples.

In contrast to logistic regression problems, estimation of item response models must use an estimate of the covariate instead of a true value of the covariate. In our context, the covariate is an estimate of the examinee's ability generated from the operational part of the on-line test. This notion is in accordance with Stocking (1990), who notes that "a sample could possibly be selected based on some available observed auxiliary information." Ford and Silvey (1980) did not solve the problem of measurement errors in the covariate when they did their study. Earlier Federov (1972) considered the development of designs in the presence measurement errors for the general linear model. Independently of researchers in the field of optimal design for logistic models, Stafanski and Carroll (1985) study the effect of measurement errors in the covariate on the asymptotic bias of the MLE of parameters of the logistic regression

model. They showed that the MLE is asymptotically biased, in other words, not consistent.

Using the ability estimate for each examinee, generated by the operational part of the on-line test, this research explores sequential D-optimal designs that are appropriate for three-parameter and other item response models. In addition this study presents an estimator of item parameters using estimates of ability in place of true ability. This research studies the relative efficiency of the normal and uniform designs. Because the designs are compared according to the Γ-optimal criterion, this research will add to the base of research that has looked at individual standard errors of estimates.

Elements of Item Response Theory

Item Response Functions

Let u_i denote a response to a single item from individual i with ability level x_i, possibly multivariate. Assume that all response variables are dichotomously scored either *correct*, $u_i = 1$, or *incorrect*, $u_i = 0$. An item response function is a function of x_i, and describes the probability of correct response of an individual with ability level x_i when presented with the item. Let $\beta^T = (\beta_0, \beta_1, ..., \beta_{M-1})$ be M parameters associated with the item. The probability of a correct response follows the form $P(x_i;\beta)$. The mean and variance of the parametric family are:

$$E\{u_i|\beta\} = P(x_i;\beta), \text{ and} \quad (1)$$

$$\sigma^2(x_i;\beta) = Var\{u_i|\beta\} = P(x_i;\beta)[1 - P(x_i;\beta)]. \quad (2)$$

An example of a family of item response functions is the celebrated family of three-parameter logistic response functions:

$$P(x_i;\beta) = \beta_2 + (1 - \beta_2)R(\beta_0 + \beta_1 x_i), \tag{3}$$

where R(z) is the logistic function:

$$R(z) = \frac{e^z}{1 + e^z}. \tag{4}$$

Another functional form for R(z) is the normal ogive, see Lord(1980) for references.

Throughout the remainder of this article, we will assume that the response variables are binary and are statistically independent given the ability level. The optimal design results of this paper can be extended to response functions other than the three-parameter family. All that is necessary is that the item response models must be differentiable in the item parameters.

MLE of Item Characteristics

The log-likelihood function of β based on independent observations $(u_i,x_i;\ i = 1,2,...,N)$ is

$$L_N(\beta;u_1,...,u_N,x_1,...,x_N) =$$
$$\sum_{i=1}^{N}\left\{u_i \log P(x_i;\beta) + (1 - u_i)\log[1 - P(x_i;\beta)]\right\}. \tag{5}$$

The maximum likelihood estimator (MLE), $\hat{\beta}$ satisfies:

$$\sum_{i=1}^{N}\frac{\left[u_i - P(x_i;\hat{\beta})\right]}{\sigma^2(x_i;\hat{\beta})}\frac{\partial P(x_i;\hat{\beta})}{\partial \beta_k} = 0,\ k = 0,...,M\text{-}1 \tag{6}$$

Under regularity conditions, the MLE is asymptotically normal with mean β and variance-covariance matrix $\mathbf{M}^{-1}(\beta)$, where $\mathbf{M}(\beta)$ is the Fisher information matrix with elements $(m_{kl}(\beta))$ equal to:

$$m_{kl}(x;\beta) = \sum_{i=1}^{N} \sigma^{-2}(x_i;\beta) \frac{\partial P(x_i;\beta)}{\partial \beta_k} \frac{\partial P(x_i;\beta)}{\partial \beta_l} \qquad (7)$$

When the ability levels are observed with error, the MLE is an asymptotically biased estimator of β, or equivalently, it is not consistent. Stefanski and Carroll (1985) suggested several modifications of the basic MLE in the logistic regression model that tend to reduce the bias. We base our modification on one of their suggestions. We first describe a plausible model for the observed ability level.

Measurement error in ability level. Let x_i denote the true ability of an examinee, and let X_i denote the observed ability obtained in an on-line testing system. Let e_i denote the measurement error associated with the on-line test. Then a measurement model is

$$X_i = x_i + e_i. \qquad (8)$$

Assume X_i, x_i, and e_i are independent and denote the variance of (X_i, x_i, e_i) by $(\sigma_{XX}, \sigma_{xx}, \sigma_{ee})$. Then

$$\sigma_{XX} = \sigma_{xx} + \sigma_{ee}. \qquad (9)$$

The reliability ratio for this measurement error model is (Fuller, 1987):

$$K_{XX} = \frac{\sigma_{xx}}{\sigma_{XX}} = 1 - \frac{\sigma_{ee}}{\sigma_{XX}}. \qquad (10)$$

We will call this ratio the *test reliability ratio*. The traditional notion of test reliability is associated with classical test theory. The definition of test reliability is relative because the variance of test scores in the population depends on the design we wish to achieve. For a fixed measurement error σ_{ee}, a test may perform well for a design where σ_{XX} is large. However, the same test with the same measurement error may perform poorly for a design where σ_{XX} is small. This notion of relative performance of test scores is well known by test practitioners; for example, an aptitude test developed for a general population is usually unsuited for the purpose of selection with a cut-off score. (This is because the measurement error of the test is not small enough to distinguish between examinees with aptitudes just to the right or left of the cut-off score.)

It is possible to control σ_{ee} in an on-line testing environment because σ_{ee} decreases as the number of administered items increases and as the on-line test administers items with high information. The reliability ratio K_{XX} will be a tuning variable for the sequential design algorithm. Values in the range 75% to 95% are meaningful. The quantity σ_{XX} depends on the sample design and determines the required value of σ_{ee} for a given value of the reliability ratio according to:

$$\sigma_{ee} = \sigma_{XX}(1 - K_{XX}). \qquad (11)$$

MLE modified for measurement error. Assume that the error terms (e_i) in model (8) are independent normally distributed with mean 0 and variance σ_{ee}. Under this measurement error model the MLEs of β and $x = (x_1, x_2, \ldots, x_N)$ maximize

$$L_N(\beta,x) = \sum_{i=1}^{N}\{u_i \log P(x_i;\beta) + (1 - u_i)\log[1 - P(x_i;\beta)] - (2\sigma_{ee})^{-1}(X_i - x_i)^2\} \quad (12)$$

The vectors $\hat{\beta}_f$, $(\hat{x}_i)$ maximizing expression (12) satisfy

$$\sum_{i=1}^{N} \frac{\left[u_i - P(\hat{x}_i;\hat{\beta}_f)\right]}{\sigma^2(x_i;\hat{\beta}_f)} \frac{\partial P(\hat{x}_i;\hat{\beta}_f)}{\partial \beta_k} = 0,\ k = 0,\ldots,M-1; \quad (13)$$

$$\hat{x}_i = X_i + \frac{\left[u_i - P(\hat{x}_i;\hat{\beta}_f)\right]}{\sigma^2(x_i;\hat{\beta}_f)} \sigma_{ee} \frac{\partial P(\hat{x}_i;\hat{\beta}_f)}{\partial x_i}, \quad (14)$$

$i = 1,\ldots,N.$

A modification of this set of equations enables an easy implementation. This modification replaces $\hat{x}_i$ by

$$\hat{x}_{i,f} = X_i + \frac{\left[u_i - P(X_i;\hat{\beta}_f)\right]}{\sigma^2(X_i;\hat{\beta}_f)} \sigma_{ee} \frac{\partial P(X_i;\hat{\beta}_f)}{\partial x_i} \quad (15)$$

We will call the resulting estimator of β the *modified* MLE. Stefanski and Carroll showed that, in logistic regression, the modified MLE reduces asymptotic bias for known σ_{ee}.

Effective sample size required for the modified MLE. Because the data follow the Bernoulli probability model, the variance of estimators based on maximum likelihood decreases approximately at the rate

$[N\bar{P}(1-\bar{P})]^{-1}$, where $\bar{P}$ is the value of the response probability averaged over the design values $x_1, x_2, \ldots, x_N$. Thus,

$$S = N\bar{P}(1-\bar{P}) \tag{16}$$

is sometimes called the *effective sample size*. For example, N = 300, 600 and $\bar{P}$ = 0.10, the effective sample size is approximately 30 and 60.

Stefanski and Carroll's Monte Carlo study showed that the increased variability of the modified estimators outweigh their savings in bias and thus under perform the regular MLE for N = 300. However, for N = 600, they showed that the modified estimator out performs the regular MLE. They chose $\bar{P}$ = 0.10 and a reliability ratio equal to 0.75 for their Monte Carlo study. Because they were interested in the performance of the estimator only, they did not investigate sequentially constructed designs and allowed the design to be observational, resulting in a normally distributed design.

Because we expect to see $\bar{P}$ fall in the range 0.4 to 0.6, and because we will use D-optimal designs instead of normal designs, the modified estimator should outperform the regular MLE for an effective sample size of no more than 60. We shall see later that, with no more than 25 per cent of the observations, a D-optimal design yields the same amount of information as does a normal design. According to the Stefanski-Carroll study, this implies that a modified MLE should start to out perform the usual MLE with an optimal design at about 0.25(60), or equivalently 15, effective observations. For $\bar{P}$ in the range 0.4 to 0.6, 15 effective observations translates into no more than N = 15/[(0.4)(0.6] < 63 actual observations.

To achieve the bias reducing properties of a modified estimator, one must control σ_{ee} with equation (11). We will see that σ_{XX} will range from 1 to about 0.4 for D-optimal designs. By equation (11), this implies that σ_{ee} will range from 0.1 to 0.25 for a fixed K_{XX} equal to 0.75. This translates into requesting the on-line test to supply estimates of ability levels with standard errors of measurement in the range 0.31 to 0.50.

This is a reasonable request to make of an on-line test that has been developed for a general population with normally distributed ability. To see this, recall that the three-parameter logistic model associates with it a population of latent ability levels, normally distributed with standard deviation $\sigma_x = 1.7$ (Lord, 1980). Assume the reliability of the on-line test is $K_{XX} = 0.97$ relative to this normal population. Using (11), it is easy to show that the variance of the measurement error is $\sigma_{ee} = (1 - K_{XX})(K_{XX})^{-1}\sigma_{xx} = (1 - 0.97)(0.97)(1.7)^2 = 0.08409$. Thus, if an on-line test has reliability 0.97 for a normal population with standard deviation 1.7, then the on-line test would have been developed to supply estimates of ability with standard errors of measurement equal to $\sqrt{0.08409} = 0.29$. The value of 0.97 for the on-line reliability is the lowest value we recommend.

Optimal Design Theory for Item Calibration

Definitions

N-point optimal designs. The sampled ability level x will be constrained to a subset $\mathcal{I}$ of R^k, where k is the dimension of the ability. $\mathcal{I}$ will be called the *design space* An N-*point design*is a collection of ability levels $x = (x_1,...,x_N)$ from the design space $\mathcal{I}$. The expected value of the i^{th}

response variable follows the response function with form $P_i(x_i;\beta)$. For each x_i denote the M-column vector of partial derivatives as $\partial P_i(x_i;\beta)/\partial\beta = (\partial P_i(x_i;\beta)/\partial\beta_0,...,\partial P_i(x_i;\beta)/\partial\beta_{M-1})$.

The design problem is to choose x with $x_i \in \mathcal{I}$, i = 1,...,N to make

$$M(x;\beta) = \sum_{i=1}^{N} \sigma^{-2}(x_i;\beta) \frac{\partial P(x_i;\beta)}{\partial\beta} \frac{\partial P(x_i;\beta)^T}{\partial\beta} \qquad (17)$$

as large as possible, where "large" is an appropriate attribute for a matrix.

Criteria for generating designs. There are several criteria for generating designs, see Silvey (1980) for an exposition. We list three criteria here. The criterion of D-optimality is the determinant of the information matrix: det $\{M(x;\beta)\}$. Its square root is inversely proportional to the volume of the confidence ellipsoid for β. The criterion of A-optimality is the trace of the inverse of the information matrix: tr $[\{M(x;\beta)\}^{-1}]$. It is proportional to the average of the variances of the estimated parameters. The criterion of *strong* optimality is the partial order on matrices induced by the condition of non-negativity: if x_1 and x_2 are two N-point designs then x_1 is better than x_2 in the *strong* sense if $M(x_1;\beta) - M(x_2;\beta)$ is non-negative definite. Both the determinant and the trace functions belong to a general class of criterion functions (ϕ) which are necessary for *strong* optimality: if x_1 and x_2 are two N-point designs such that x_1 is better than x_2 in the *strong* sense such that $M(x_1;\beta) - M(x_2;\beta)$ is non-negative definite then $\phi\{M(x_1;\beta)\} \geq \phi\{M(x_2;\beta)\}$.

We have introduced the criterion of A-optimality primarily because of its obvious relation to the studies of Wingersky and Lord (1984) and Stocking (1990). These studies indicate that, when using a random design

with random sampling, a rectangular distribution over ability level is better than a normal one according to the criterion of A-optimality. We will concentrate on the criterion of D-optimality.

The solution to the problem of finding an N-point D-optimal design reduces to solving the mathematical programming model:

$$\text{Maximize } \det\{M(x;\beta)\}$$

$$\text{such that}$$

$$x_i \in \mathcal{I},\ i = 1,\ldots,N. \tag{18}$$

This mathematical programming model has wide applicability to item response theory. All that one needs is the Fisher information function (7) for the targeted item response models. It can also be expanded to cover designs for simultaneous item calibration.

Approximate theory of D-optimal designs. Many theoretical techniques are available in D-optimal design theory for the problem with the criterion extended to probability distributions over the design space. This is called the *approximate* theory of optimal design. We derive approximate D-optimal designs for the purpose of comparing sequential designs, normally distributed designs, and rectangular designs.

Let us extend the criterion to probability distributions over the design space by first considering a finite design space. Denote points of the design spaces by the distinct values $x_{(1)},\ldots,x_{(r)}$. A design will replicate these design sites $n_1,\ldots,n_r$ times, respectively. We associate with an N-point design x a probability distribution on $\mathcal{I}$: η_N, which puts probability $p_i = n_i/N$ at $x_{(i)}$. Let x be a random variable with distribution η_N and redefine the information matrix associated with the η_N as:

$$M(\eta_N;\beta) = E\left[\sigma^{-2}(x;\beta)\frac{\partial P(x;\beta)}{\partial\beta}\frac{\partial P(x;\beta)}{\partial\beta}^T\right] \qquad (19)$$

$$= \sum_{i=1}^{r} p_i \sigma^{-2}(x_i;\beta)\frac{\partial P(x_i;\beta)}{\partial\beta}\frac{\partial P(x_i;\beta)}{\partial\beta}^T = N^{-1}M(x;\beta).$$

Now it is straight forward to define the information matrix of an arbitrary probability distribution over a design space, as follows. Assume the usual probability space over $\mathcal{I}$. Let η be a probability distribution and x a random ability level with distribution η, define

$$M(\eta;\beta) = E\left[\sigma^{-2}(x;\beta)\frac{\partial P(x;\beta)}{\partial\beta}\frac{\partial P(x;\beta)}{\partial\beta}^T\right] \qquad (20)$$

Duality. Sibson (1972) showed that the approximate D-optimal problem is dual to another mathematical programming model. Sibson's result deals with the linear model. We extend this result to non-linear models by defining the manifold in R^M induced by the design space $\mathcal{I}$ as follows:

$$\mathcal{M} = \{z \in R^M: z = [\sigma(x;\beta)]^{-1}\partial P(x;\beta)/\partial\beta, x \in \mathcal{I}\}. \qquad (21)$$

The set $\mathcal{M}$ depends on β. Sibson' result shows that the D-optimal problem is dual to the problem of finding a minimal content ellipsoid contained in R^M centered at the origin that contains the manifold $\mathcal{M}$. We present this fact not because it leads to practical solution methods but because it leads to deep theoretical insight about the nature of optimal designs.

Relative efficiency. To evaluate the efficacy of a given design we introduce the notion of *relative efficiency*. Let $M^*(\beta)$ be the information

matrix of the approximate optimal design for a given value of the item vector, β. Then the relative efficiency of a design, η, is defined as

$$\text{eff}(\eta) = \frac{\det\{M(\eta;\beta)\}}{\det\{M^*(\beta)\}}. \qquad (22)$$

Random-seeding and uniform designs. An observational design would occur if the experimenter did nothing and let the observations occur with random ability levels. For example, practitioners sometimes assume that the naturally occurring distribution of a unidimensional ability is normal. The normal distribution associated with the three-parameter logistic model has a mean of zero and a standard deviation of 1.7. Let ϕ denote the standard normal distribution. For this example $\eta(x) = \phi(x/1.7)$. This design occurs naturally if examinees are selected at random to receive an experimental item. This is known as *random-seeding* of experimental items. The rectangular or uniform design studied by Wingersky and Lord (1984) and Stocking (1990) also can be formulated with an η.

Algorithms for D-optimal designs. Using the two-parameter logistic model, Ford (1976) obtained approximate D-optimal designs that have closed formulas, see Appendix A. In general, no closed solutions to model (18) exit. We will use Ford's approximate solutions to compare with other designs. We will employ a search method to solve the problem (18) for N-point designs when N is small, which we will use to find sequentially constructed designs. We base our method on those that have been developed for the linear model, which we discuss now.

In case of linear statistical models, there exit several heuristic search techniques (Federov p. 167, 1972; Welch,1982; and Haines, 1987). Federov's algorithm is based on gradient search, but applies to only linear

models and finds approximate, not N-point, optimal designs. Haines' algorithm is based on simulated annealing, handles both D-optimal and A-optimal criteria, produces N-point designs, and is simple to code. Welch's algorithm is based on branch-and-bound search over designs confined to a finite set of pre-specified points and finds exact N-point D-optimal designs. Among these three algorithms, branch-and-bound search requires the least amount of computer time for most practical problems. We employ the branch-and-bound heuristic to find small-sample designs for use in the sequential design scheme described in the next section. We present the branch-and-bound search in Appendix B.

Results and Conclusion

Approximate optimal designs for two-parameter logistic items. Ford (1976) derived a formula for approximate D-optimal designs for the two-parameter logistic model. We list the description in Appendix A for the design space $\mathcal{I} = [-1, +1]$ and we will use this design space for the following results. This design space corresponds to the notion of concentrating the design to single out individuals with more informative ability levels than would be the case with a normal distribution (Stocking, 1990).

The main characteristic of D-optimal designs is that the approximate optimal design puts one-half its probability at each of *two* points. This follows from the duality theory and the shape of the manifold in R^2, induced by the design space. To see this, note first that, for the two-parameter logistic model M=2,

$$[\sigma(x;\beta)]^{-1}P'(x;\beta) = \{P(x;\beta)[1-P(x;\beta)]\}^{1/2}, \qquad (23)$$

hence

$$\mathfrak{M} = \{z \in R^2: z = \{P(x;\beta)[1-P(x;\beta)]\}^{1/2}[1,x]^T, x \in \mathfrak{X}\} \quad (24)$$

The manifold $\mathfrak{M}$ is a smooth curve in R^2 because the logistic model is infinitely differentiable. Graphically, it resembles the curve of a smooth "C" with ends curved back toward the origin. Analytically, the form of its defining equations are not elliptical. Hence, the optimal design has a two-point distribution because the subset S touches a minimal area ellipse in R^2 centered at the origin at two points .

In Table 1, we have listed the a and b parameters of three logistic curves along with their support points of the approximate D-optimal design . In terms of our parametric representation, the conventional discrimination parameter, a, and difficulty parameter, b, of the two-parameter logistic model are:

$$a = \beta_1 \text{ and } b = -\frac{\beta_0}{\beta_1}. \quad (25)$$

Each set of parameters are representatives from the regions B_1, B_2 and B_3 (A-1 to A-3). We will characterize these three sets of item parameters as "low," "medium," and "high" in reference to the value of the discrimination parameter. The three values of the discrimination parameter appear to reflect the values seen in practice. Note that the support points of the optimal design are not found at the extreme points, but at the interior points of [-1, +1] in some cases.

Table 1 about here.

In these two-parameter logistic models, we have restricted the design space to the interval [-1, +1]. If we widened the interval, the support points for the "low" and "medium" items will no longer be the end points of the interval. Thus optimal designs are not merely the extreme points of the interval. This is the major feature in which optimal designs for non-linear models are different from optimal designs of linear models. The reason for this phenomenon is that manifold $\mathcal{M}$, defined above, is non-linear in x.

For the three-parameter logistic model, the manifold $\mathcal{M}$ is a subset of R^3. It does not have an elliptical surface and therefore it will touch its minimum content ellipsoid in three points. The value of these three points and the corresponding design probabilities are complicated and will not be presented here.

We will not use these theoretical results in our construction of sequential designs because we desire to obtain designs for response models other than the two or three parameter logistic models. Also we seek exact N-point optimal designs for which approximate optimal designs may not fit well. However, we will use these theoretical results to compare with some ad hoc designs, such as the normal or rectangular distribution of ability level.

Relative efficiencies of some *random-seeding* designs for estimating the two-parameter logistic response model. The relative efficiency of a continuous design η for a two-parameter logistic model is obtained from:

$$M(\eta;\beta) = \int_{\mathcal{I}} P(x;\beta)[1-P(x;\beta)]\begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}\eta(x)dx. \qquad (26)$$

In case η corresponds to a rectangular or normal density, the integrals may be readily evaluated with quadrature methods.

We have determined the relative efficiencies for these two designs for the three logistic models discussed previously (Table 2). Although the normal design performs better than the uniform design, we note that both the normal and uniform designs perform poorly. Indeed the best relative efficiency that the normal design attains is only 23.56% for the item parameters (2.3851, -.1745). In estimating these item parameters, this means roughly that for 100 observations from the normal design, one may obtain the same amount of accuracy with 12 observations at -0.82 and 12 observations at 0.47.

Table 2 about here.

That the normal design performs better than the rectangular design raises a question about the consistency of these results with those of Wingersky and Lord (1984) and Stocking (1990). An explanation is that we are comparing designs according to the criterion of D-optimality whereas they compare designs by a criterion directly related to A-optimality. Rectangular designs are better (worse) than normal designs according to the criterion of A-optimality (D-optimality). This result is not inconsistent with the criterion of strong optimality because both D-

optimality and A-optimality criteria are necessary, but not sufficient for a design to be strongly optimal. We have not investigated the criterion of strong optimality for rectangular and normal designs.

Sequential Design Theory and Monte Carlo Study

Methods

Sequential designs. Because of the dependence of an optimal design on the item parameters, it is impossible to employ the optimal design in practice. To overcome this drawback, we sequentially construct reasonable designs. The construction procedure collects the total sample in small subdesigns, size n, that are n-point D-optimal for the current estimates of the item parameters. All this takes place within the environment of on-line tests, which provide estimates of ability levels. The estimates of the item parameters gradually improve as the overall sample size, N, increases. To ensure this improvement, the estimate accounts for error in the estimated ability levels using the modified MLE (13) and (15).

We consider a simple framework for sequentially constructing a design. We obtain a sequential design by repeatedly cycling through three steps (Figure 1). Step one obtains responses to the item from a small number of examinees whose abilities satisfy the design from step three. Step one accumulates these data with prior data and obtains estimates of the item parameters. Step three obtains a small sample design that is an exact n-point optimal design for model (18), where the estimates of step two substitute for β. A sequential design results in an empirical design that will rarely equal an exact n-point D-optimal design; but will, however, approximate the optimal design associated with the unknown item

parameters. We determine how well the sequential designs perform by computing their relative efficiencies (22).

Figure 1 about here.

Simulation study. We fully describe the sequential algorithm in Appendix C. One must fix several tuning parameters for each application of the algorithm. These tuning parameters are: n, the size of the subdesign; N, the overall sample size or stopping time; $K_{XX'}$, the test reliability (10); S, the effective sample size (16). We obtained simulations of designs for calibrating two-parameter logistic items. We varied the tuning parameters as follows: n = 3, 5, 15; N = 200, 400; $K_{XX'}$ = 0.75, 0.80, 0.85, 0.90, 0.95, 1.00; S = 1, 15, 30, ∞. For brevity, we did not duplicate some of the reliability ratios between N = 200 and 400.

We fixed the design space to be twenty points unequally spaced along the interval [-1, +1] (Table C-1). These twenty points correspond to ten pairs of support points for ten approximate D-optimal designs. The items, for which these ten designs are optimal, are listed in Table C-1 under the column heading "associated items." Assuming that experimental items have difficulties between -1 and +1, this design space enables the algorithm to expose items to examinees with ability levels no farther away than two units from the difficulty parameter. We chose these ten items to be representative of the range of discrimination parameters and the three regions B_1, B_2, and B_3 defined in the equations (A-1, A-2, and A-3).

By symmetry arguments, the simulations results apply to items with difficulty parameters obtained from the difficulty parameters of the items in Table 1 reflected around zero. The design space would be obtained also by reflecting the design space in Table C-1 around zero.

Results and Conclusion

The relative efficiencies of sequentially constructed designs are presented in Tables 4 and 5. The overall impression conveyed by the tables is that sequentially constructed designs are reasonable designs for item calibration. Indeed, the lowest efficiency is 0.77 for N ~ 200, and 0.72 for N ~ 400.

Performance of modified MLE's. Generally speaking, the regular MLE was superior to the modified one. The exceptions were for the large discrimination parameter in two cases: a) with n = 15, N = 180, in Table 3; and b) with n = 3, N = 402, in Table 4.

Large reliability ratios tend to degrade the performance of the MLE: the more error in the estimated ability level, the worse the MLE performs. Even when there is no error in the ability level, the modified MLE does worse than the unmodified MLE for n = 15 and the item with the large discrimination parameter (Table 3). This is because the performance of any design, that is not optimal, is poor; and, the instability of the modified MLE exacerbates this problem, especially when any one subsample represents a high proportion of the overall design.

Effects of sample size. It appears that n = 5 is superior to 3 in Tables 3 and 4. However too large a subsample size is bad, as seen with n = 15 in Table 3. Because our simulation ran too long, we do not report results for n = 15 and N = 400. As expected, the greater the overall sample size,

the higher relative efficiency attained with the sequentially constructed designs.

It was thought that the modified MLE would compensate better than it did for the error in the ability level for the effective sample sizes S, and overall sample size N, that we studied. However, asymptotic theory suggests that the modification becomes worthwhile at a larger sample size than N = 400. Another possible reason for these poor results is that the effective sample sizes studied were too small. In Table 4, we see that the modified and regular maximum estimators are operatively equal for S = 30. We do not present the results here but we have studied S = 40, 50, ..., 100. We found that modified MLE does only slightly better than the regular MLE.

Laying the performance of the modification aside, these results compare extremely well with random-seeding of items. We saw that the best that random-seeding did was 27% efficiency, whereas, the worst that sequential designs did is about 95% efficiency with the best configuration: no modification, n = 5, and N = 400 (Table 4).

In conclusion, these results suggest that one should implement sequentially constructed designs utilizing the regular MLE, a subsample size of 5, and overall sample size of N = 400. This configuration performs well even for items with large discrimination parameters and on-line tests with low reliability.

Discussion

Modified MLE. Let us assume that we have a consistent estimator of β. The relative efficiency of a sequential design approaches unity as the

total sample size increases, or in other words, a sequential design is asymptotically optimal (Wu, 1985). However, the rate at which the relative efficiency approaches one is important. As we have seen, an instable estimator results in inefficient designs for moderate sample sizes. The theory on this topic falls in the general area of second order efficiency; however, the theory is incomplete for sequential designs.

Designs for simultaneous item estimation. Practical constraints may dictate that designs for several items shall be constructed where items must "compete" with each other for a *limited* number of examinees with optimal ability levels. In addition, frugal utilization of each calibration session requires that all the examinees be used. The mathematical programming model for this situation is as follows: Let $\mathcal{X} = (x_i: i = 1,...,r)$ be a finite collection of candidate ability levels; m_i denote the number of examinees with ability x_i available; and $m_+ = n$. The "plus" notation denotes summation over the index. Suppose we are to calibrate c items, each with response function $P_j(x;\beta)$, $j = 1,...,c$. A collection of c, n_j-point, designs for simultaneously calibrating c items is (n_{ij}), non-negative integers, where $n_{+j} = n_j$, $j = 1,...,c$; $n_+ = n$. Associate with each n_j-point design a probability distribution on $\mathcal{X}$: η_{nj}, which puts probability $p_{ij} = n_{ij}/n_j$ at x_i. Let $M_j(\eta_{nj};\beta)$ denote the information matrix associated with item j. A mathematical programming model is:

$$\max \sum_{j=1}^{c} \frac{n_j}{n} \log \det\{M_j(\eta_{nj};\beta)\} \quad (27)$$

such that

$$n_{i+} = m_i,\ i = 1,\ldots,r; \quad (28)$$

$$n_{+j} = n_j,\ j = 1,\ldots,c; \quad (29)$$

$$m_+ = n_+ = n \quad (30)$$

$$n_{ij} \geq 0,\ \text{integer}. \quad (31)$$

The criterion (27) is related to, but not equal to, D-optimality; that is to say the criterion does not correspond to the joint confidence ellipsoid of all c sets of item parameters. Another criterion is the simple summation of all c information matrices; however, this criterion is not equal to the criterion of D-optimality. Because the "log-det" function is strictly concave over the space of non-negative definite matrices, the criterion (27) is a lower bound to the logarithm of the weighted sum of c information matrices, with weights n_j/n.

The solution to the problem (27) - (31) may be solved with the branch-and-bound technique. The values for m_i in constraint (28) are uncontrolled, resulting from expected flows of examinees at individual testing sites. Constraint (29) enables the practitioner to control the proportion of the total observations allocated to any one item. This is an important degree of freedom as one may desire to spend less observations on items with imprecise estimates of difficulty relative to the other items. Constraint (30) is also uncontrolled and determined by the flow of examinees.

On-line item seeding and non-interference with the testing of the subjects' ability level. To limit the exposure of examinees to inappropriate

items, one may either constrain the design space, add more constraints to the mathematical programming problem, or reformulate the information matrix. Our simulations limited the design space to the interval [-1, +1]. If most experimental items have difficulty in the range -1 to +1, then this design space makes it unlikely that examinees are exposed to an item more than two units away from their ability. Other intervals could also be used.

If a range of item difficulties, wider than [-1, +1], is anticipated, then one could allow the candidate design points to be spread out in the appropriate interval, and also place constraints on the distance between the item difficulty and the candidate design points. We propose the constraint on each design point: $\{P(x_i;\beta)[1 - P(x_i;\beta)]\}^h \geq v$, where v and h are non-negative fixed constants. The effect of this constraint is to eliminate certain candidate design points that are outlying relative to the item difficulty.

Another approach is to modify the Fisher information matrix so that the D-optimal design points are not the extreme points in the design space. An easy modification is:

$$\mathbf{M}(x;\beta,h) = \sum_{i=1}^{r} \sigma^{-2+h}(x_i;\beta) \frac{\partial P(x_i;\beta)}{\partial \beta} \frac{\partial P(x_i;\beta)^T}{\partial \beta}, \qquad (32)$$

where h is a non-negative constant. The integer programming problems using this criterion is solvable with the methods proposed here.

References

Donev, A. N. and Atkinson, A. C. (1988). An adjustment algorithm for the construction of exact D-optimum experimental designs. *Technometrics, 30*, 429-433.

Federov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Ford, I. (1976). Ph.D. Thesis, University of Glasgow.

Ford, I., & Silvey, S. D. (1980). A sequentially constructed design for estimating a non-linear parametric function. *Biometrika, 67*, 000-000.

Haines, Linda M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models. *Technometrics, 29*, 439-447.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Erlbaum.

Lord, F.M. (1971). Tailored testing, an application of stochastic approximation. *Journal American Statistical Association, 66*, 707-711.

Mitchell, T. J. (1974). An algorithm for the construction of D-optimal experimental designs. *Technometrics, 16*, 203-210.

Silvey, S. D. (1980). *Optimal design*. New York: Chapman and Hall.

Stefanski, L. A. and R. J. Carroll. (1985). Covariate measurement error in logistic regression. *Annals Statistics, 13*, 1335-1351.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461-475.

Welch, W. J. (1982). Branch-and-bound search for experimental designs based on D optimality and other criteria. *Technometrics, 24*, 41-48.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Wu, C.F.J. (1985). Asymptotic inference from sequential design in a nonlinear situation. *Biometrika, 72*, 3, 553-558.

Appendix A

Approximate Optimal Designs for the 2PL Model

Suppose we let the design space $\mathcal{X}$ be [-1, +1] and it is known that $\beta_0 > 0, \beta_1 > 0$. Ford (1976) has shown that for the two parameter logistic function the approximate optimal design puts one-half the mass at one point and one-half at another point. These two points depend on the value of the item parameter vector, β in the following way.

Let c be the positive solution of the equation $e^z = \frac{z+1}{z-1}$, $c \sim 1.5434$. Also let

$$B_1 = \{\beta: \beta_0 > 0, \beta_1 > 0, \beta_1 - \beta_0 \geq c\} \quad \text{(A-1)}$$

$$B_2 = \{\beta: \beta_0 > 0, \beta_1 > 0, \beta_1 - \beta_0 < c, \exp(\beta_0 + \beta_1) > \frac{\beta_1 + 1}{\beta_1 - 1}\} \quad \text{(A-2)}$$

$$B_3 = \{\beta: \beta_0 > 0, \beta_1 > 0, \beta_1 - \beta_0 < c, \exp(\beta_0 + \beta_1) \leq \frac{\beta_1 + 1}{\beta_1 - 1}\} \quad \text{(A-3)}$$

Then

(i) if $\beta \in B_1$, the support points are $\frac{c - \beta_0}{\beta_1}, \frac{-c - \beta_0}{\beta_1}$, (A-4)

(ii) if $\beta \in B_2$, they are -1 and x_u where x_u is the solution of $\exp(\beta_0 + \beta_1 x) = \frac{2 + (x+1)\beta_1}{-2 + (x+1)\beta_1}$, (A-5)

(iii) if $\beta \in B_3$, they are -1 and +1. (A-6)

This rather complicated design can be roughly summarized by saying the more steep the response curve is, the more the support points are squeezed to the center of the design space; the more flat, and hence linear,

the response curve is, the more the support points are pushed to the extremes of the design space.

Appendix B

The Branch-and-Bound Procedure

Let x_i, i = 1,2,...,r denote candidate design points. The N-point D-optimal design problem is to place N observations at the r design points so as to maximize $\det\{M(\eta_N;\beta)\}$. There are $\begin{bmatrix} r+N-1 \\ N \end{bmatrix}$ possible designs, some leading to a zero determinate. Instead of performing an exhaustive search over all possible designs, it is possible to partition the set of designs and to perform searches over a much smaller set of designs. Let $l = (l_1,l_2,...,l_r)$ and $u = (u_1,u_2,...,u_r)$ be collections of non-negative integers less than or equal to N. The maximum determinate exceeds or equals the solution to:

$$\text{maximize } \det\{M(\eta_N;\beta)\}$$

$$\text{such that } \sum_{i=1}^{r} n_i = N,$$

$$l_i \le n_i \le u_i,\ i = 1,2,...,r. \qquad \text{(B-1)}$$

We call this maximization a *node*. The original maximization is called the *root node* with all $l_i = 0$ and $u_i = N$. The collection (B-1) of designs is further subdivided into two nonempty partitions, as follows:

$$n_j = l_j,\ l_i \le n_i \le u_i,\ i \ne j \qquad \text{(B-2)}$$

$$l_j + 1 \le n_j \le u_j,\ l_i \le n_i \le u_i,\ i \ne j \qquad \text{(B-3)}$$

where the way j will be chosen is described shortly. Thus we create two nodes by replacing (B-1) with either (B-2) or (B-3).

Thus every node has either zero or two branches leading from it, creating a binary tree with the $\begin{bmatrix} r+N-1 \\ N \end{bmatrix}$ N-point designs located at the extreme nodes. We guide the search for the optimal design by going up the tree along branches that are not suboptimal.

We avoid suboptimal branches by calculating a bound on $\det\{M(\eta_N;\beta)\}$ over all designs leading from a common node (l,u) on the branch as follows. Define

$$N' = \sum_{i=1}^{r} l_i \tag{B-4}$$

$$M(l,\varepsilon;\beta) = \frac{1}{N}\sum_{i=1}^{r} (l_i + \varepsilon)\sigma^{-2}(x_i,\beta_i) \frac{\partial P(x_i,\beta_i)}{\partial \beta} \frac{\partial P(x_i,\beta_i)^T}{\partial \beta} \tag{B-5}$$

$$d(x_i,l,\varepsilon;\beta) = \sigma^{-2}(x_i,\beta_i) \frac{\partial P(x_i,\beta_i)^T}{\partial \beta} M^{-1}(l,\varepsilon;\beta) \frac{\partial P(x_i,\beta_i)}{\partial \beta} \tag{B-6}$$

where ε is a small positive number to ensure that $M(l,\varepsilon;\beta)$ is positive definite.

The determinant of the information matrix where the design satisfies the constraints (B-1) satisfies the following bound (Welch, 1982)

$$\det\{M(\eta_N;\beta)\} \le \det\{M(l,\varepsilon;\beta)\}\{1 + \bar{d}(l,\varepsilon;\beta)\}^{N-N'} \tag{B-7}$$

where

$$\bar{d}(l,\varepsilon;\beta) = \max_{\substack{i=1 \text{ to } r \\ l_i < u_i}} d(x_i,l,\varepsilon;\beta) \tag{B-8}$$

The value for j that is used to branch at the node is the value of i for which the maximum $d(x_i,l,\varepsilon;\beta)$ is attained.

Appendix C

The Sequential Design Algorithm

Step 0

a) Choose S, the effective sample size for switching from the regular MLE to the modified estimator.

b) Choose N_0, the maximum number of observations to be gathered via the sequential design algorithm.

c) Choose $K_{XX'}$ the test reliability ratio.

d) If this is a simulation choose the item response model and its parameter vector β.

e) Choose n, the sample size for the subdesign. Set N = n initially.

f) Choose X_i, i= 1,...,n, an initial design.

Step 1

a) Pool the design points X_i, i = 1,...,n with all prior design points.

b) Determine σ_{XX} for the set of pooled design points.

c) Set $\sigma_{ee} = \sigma_{XX}(1 - K_{XX'})$.

d) If this is a simulation, randomly generate $x_i = X_i + e_i$, where e_i follows $N(0,\sigma_{ee})$, i = 1,...,n. If this is part of a real time system, find n examinees with estimated ability level X_i and measurement error σ_{ee}, i = 1,...,n.

e) If this is a simulation, obtain n random responses u_i at latent ability x_i, according to the true item response model $Pr\{u_i = 1|x_i,\beta\}$. If this is a part of a real time system, obtain a response to the item from each of the chosen n examinees.

f) Pool u_i, X_i with all prior data, i = 1,...,n.

Step 2

a) Find $\bar{P}$ for the current design, $\bar{P} = N^{-1} \sum_{\text{all data}} u_i$.

b) Find the effective sample size, $N' = N\bar{P}(1-\bar{P})$.

c) If N' > S, obtain modified maximum likelihood estimates of the item parameters. Otherwise obtain regular maximum likelihood estimates of the item parameters.

Step 3

a) N = N + n

b) If $N \geq N_0$, stop. Otherwise continue.

c) Based on the current item parameter estimates, find $X_1,...,X_n$, the exact n-point optimal design using branch-and-bound and the criterion of D-optimality.

d) Go to step (1.a).

Table 1

Optimal Designs on [-1, +1] for Two-Parameter Logistic Items

	Two-Parameter Logistic Response Function					
	Low		Medium		High	
a, b	1.2030, -.3459		1.7326, -.2402		2.3851, -.1745	
x	-1	1	-1	.73	-.82	.47
$\eta(x)$	.5	.5	.5	.5	.5	.5
μ_x	0		-.135		-.175	
σ_{xx}	1		.748		.416	

Table 2

Relative Efficiencies for a Normal N(0, 1.7) and a Rectangular U[-1, +1] Design in Estimating Two-Parameter Logistic Items

	Two-Parameter Logistic Response Function (a,b)		
	Low	Medium	High
Design	1.2030, -.3459	1.7326, -.2402	2.3851, -.1745
N(0, 1.7)	.2028	.2252	.2356
U(-1, +1)	.0859	.1194	.1544

Table 3

Relative Efficiencies of Sequential Designs N ~ 200

	Magnitude of Discrimination Parameter					
	Small		Medium		Large	
	Effective Sample Size§					
Reliability	1	∞	1	∞	1	∞
	n = 3, N =198					
.85	.96	.98	.90	.95	.89	.86
.90	.98	.99	.87	.95	.79	.87
.95	.98	.99	.93	.96	.80	.88
1.0	.99	.99	.93	.93	.89	.89
	n = 5, N =200					
.85	.95	.98	.88	.97	.89	.90
.90	.95	.98	.86	.97	.88	.91
.95	.96	.98	.96	.95	.90	.91
1.0	.98	.98	.95	.95	.91	.91
	n = 15, N =180					
.85	.99	1.0	.94	.95	.86	.80
.90	1.0	1.0	.90	.93	.87	.77
.95	1.0	1.0	.91	.93	.85	.77
1.0	1.0	1.0	.93	.93	.85	.77

§"∞" means the effective sample size was so large that the regular M.L.E. was used always.

Table 4

Relative Efficiencies of Sequential Designs N = 400

	Magnitude of Discrimination Parameter											
	Small				Medium				Large			
	Effective Sample Size§											
Rel	1	15	30	∞	1	15	30	∞	1	15	30	∞
	n = 3, N = 402											
.75	.95	.96	.99	.99	.73	.95	.94	.97	.91	.91	.93	.89
.80	.97	.96	.98	.98	.72	.95	.95	.97	.92	.94	.95	.92
.85	.98	.95	.98	.98	.85	.96	.96	.97	.93	.93	.94	.91
	n = 5, N =400											
.75	.82	.97	.98	.99	.77	.97	.98	.98	.91	.91	.94	.95
.80	.92	.99	.99	.99	.84	.97	.97	.98	.89	.89	.94	.95
.85	.97	.97	.99	.99	.93	.97	.98	.98	.90	.90	.94	.95

§"∞" means the effective sample size was so large that the regular M.L.E. was used always.

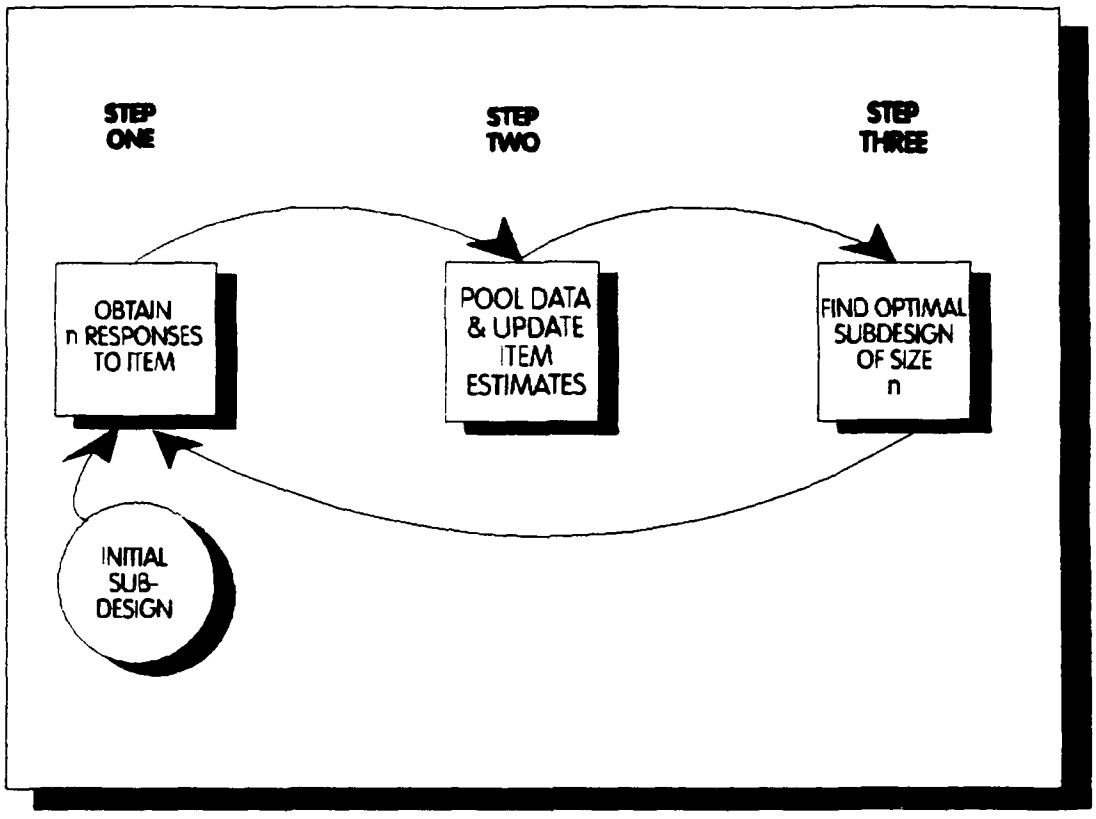
Table C-1

Candidate Ability Levels

	Support Points		Associated Items	
Pair Number	Left	Right	a	b
1	-1.000	1.0000	1.2030	-.3459
2	-.9325	.9925	1.3922	-.2989
3	-.8735	.8529	1.5056	-.2764
4	-.8216	.7315	1.6191	-.2570
5	-.7755	.6371	1.7326	-.2402
6	-.7343	.5753	1.8461	-.2254
7	-.6972	.5364	1.9595	-.2124
8	-.6637	.5025	2.1014	-.1980
9	-.6333	.4726	2.2432	-.1855
10	-.6055	.4461	2.3851	-.1745

Figure Caption

Figure 1. Flow of tasks for sequentially constructing a design for non-linear response models.


STEP ONE
STEP TWO
STEP THREE
OBTAIN n RESPONSES TO ITEM
POOL DATA & UPDATE ITEM ESTIMATES
FIND OPTIMAL SUBDESIGN OF SIZE n
INITIAL SUB-DESIGN